Genomics Data Warehousing Forum
January 19, 1999
Sponsored by
Center for Applied Scientific Computing (CASC)
Lawrence Livermore National Laboratory

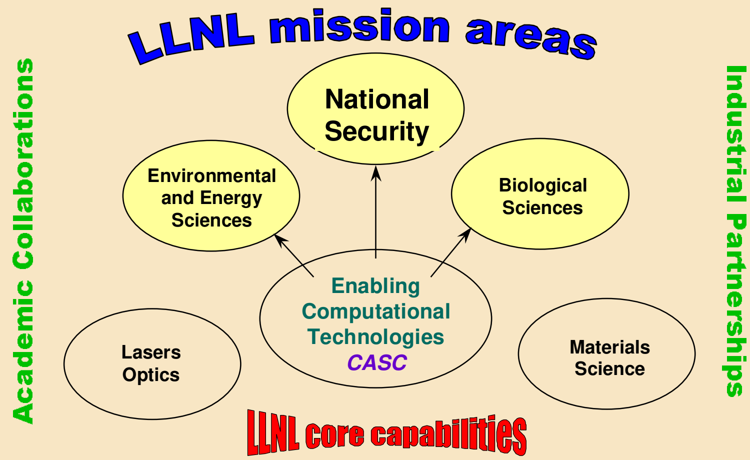
Meeting Agenda

10:30	**Welcome and Forum Overview**	**Terence Critchlow**
	Collaborating with LLNL in Life Sciences Informatics	**Steve Ashby**
	Life Science Informatics Overview	**Ron Musick**
	DataFoundry Overview	**Terence Critchlow**
11:45	***Lunch and Informal Discussions***	
1:00	**DataFoundry's Meta-Data Infrastructure**	**Terence Critchlow**
	Data Analysis in Scientific Warehouses	**Madhavan Ganesh**
	The DataFoundry Architecture	**Ron Musick**
	Joint Genome Institutes Perspective on Genome Data Warehousing	**Tom Slezak**
2:00	**Summary**	**Terence Critchlow**
2:15	**Adjourn**	

Collaborating with LLNL in Life Sciences Informatics

Steven Ashby
Director
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory

LLNL is a multi-programmatic, multi-disciplinary national security lab



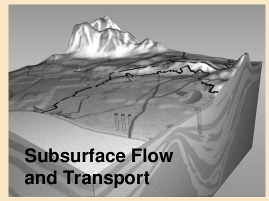
CASC is the Laboratory's center for computer science and mathematics R&D

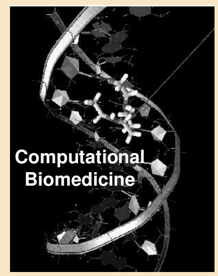
- Partner with academia and industry to research, develop, and deploy enabling computational technologies for large-scale scientific simulation
- Partner with DOE programs to conduct collaborative scientific investigations

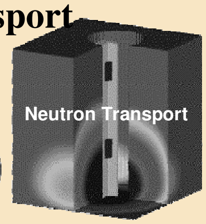
LLNL has a long history of forming multi-disciplinary teams to solve challenging problems of national importance

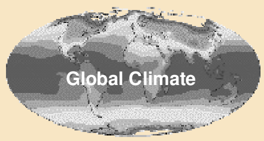
CASC expertise is being applied to a variety of application areas

- Data management
- Pattern recognition
- Scientific visualization
- Adaptive mesh refinement
- Scalable linear solvers
- Radiation & neutron transport









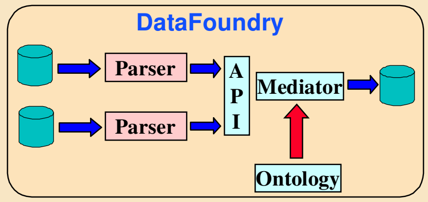
The Lab has identified bioinformatics as an important technology to pursue

> "LLNL is heavily involved in biology and biotechnology research. Providing the information infrastructure required to support tasks in these areas as well as other scientific activity at LLNL is an important focus of our research and development."
>
> -Bruce Tarter
> Laboratory Director

- Partner in DOE's Joint Genome Institute
- Chromosome 19 sequencing
- Protein Structure Prediction Center
- and...

DataFoundry is a data warehouse for bioinformatics

- Ongoing research project
- Unique architecture and meta-data infrastructure



- Integration and querying of multiple, dynamic scientific data sources

Today's forum will discuss our current status and future research plans

DataFoundry team is multidisciplinary

- Computer Science
  - Ron Musick — *meta-data, data mining*
  - Terence Critchlow — *meta-data, database integration*
  - Madhaven Ganesh — *data mining, dbase integration*
- Domain expertise
  - Krzysztof Fidelis — *structural biologist*
  - Tom Slezak — *bioinformatics (20+ years)*
  - Adam Zemla — *structural biologist*

We are actively seeking industrial partners to team with us on future research and development

Several ways to partner with LLNL in a response to Life Sciences Informatics RFP

- License agreement
  - non-exclusive use of LLNL developed technology
- CRADA
  - collaborate on a joint research project
  - may negotiate exclusive rights to technology developed by LLNL under CRADA
- Joint proposals
  - rights depend on grant

Life Science Informatics: An Opportunity for Collaboration

Ron Musick

Terence Critchlow *Madhavan Ganesh*

Center for Applied Scientific Computing

Krzysztof Fidelis *Tom Slezak*

Biology and Biotechnology Research Program

Lawrence Livermore National Laboratory

LSI is a University of California matching grant program

- **Program sponsor:**
  - **Industry-University Cooperative Research program**
- **Task Force:**
  - **Representatives from 9 UC campuses, 3 National Labs**
- **Program:**
  - **LSI has $2 million for matching grants in FY99**
  - **Funding available for up to $4 million starting FY00**
  - **Grants match $1-$1 private sponsor funding**
  - **Typical grants: research oriented, 1-2 years, $50,000 to $450,000 LSI funding**

http://uc-industry.berkeley.edu/IUCRP-Partnerships/LSI/LSI_RFP.htm

CASC

Goals of the Life Science Informatics program

- **Foster outstanding *early stage research*, with emphasis on novel, multidisciplinary approaches**
- **Support education and the research training environment**
- **Speed technology transfer to California industry for the benefit of Californians**
- **Assess and communicate the economic impact of research and education in life science informatics**

CASC

Specifics of the grant process

- Intellectual property
  - Industrial partner can negotiate an exclusive commercial license to patentable inventions
- Proposal issues
  - LSI is looking for large proposals
- Dates and deadlines
  - 12/23/98 CFP came out
  - 2/1/99 Non-binding letter of intent
  - 3/5/99 Proposals & binding letters of intent due
  - 3/99 Peer review, 4/99 notification
  - 6/14/99 Research agreement must be in place
  - 6/15/99 Funding to begin

Examples of research likely to be supported

- Linking databases containing disparate information
- Comparison of plant and human genomes
- Linking computational models across scales
- Innovative software architectures and analytic capabilities for health care computing systems
- New technologies to speed research results into clinical trials and practice

Very broad -
But must avoid DiMI, BioSTaR

Matching funds make the LSI program very attractive

- LSI matching funds benefits
  - Instant leveraging of research investment
  - CA (24%) & Federal R&D tax credits for investment in academic research
  - Opportunity to negotiate intellectual property rights
- Working with DataFoundry
  - Leveraging the research of a multidisciplinary team
  - Access to the DataFoundry technology
  - Access to LLNL's outstanding facilities and research

DataFoundry

Terence Critchlow

Ron Musick *Madhavan Ganesh*

Center for Applied Scientific Computing

Krzysztof Fidelis *Tom Slezak*

Biology and Biotechnology Research Program

Lawrence Livermore National Laboratory

Outline

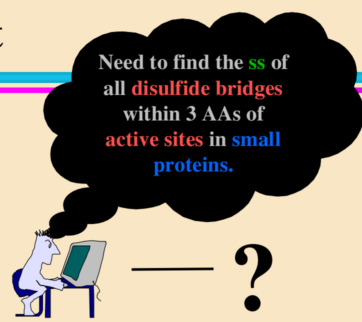
The capability of bioinformatics tools to access data from several different gene and protein databases will be the top industry challenge in 5 years.

Frost & Sullivan: US Bioinformatics in the Emerging Drug Discovery Marketplace 1998

- **Motivation**
- **Demo**
- **Conclusions**

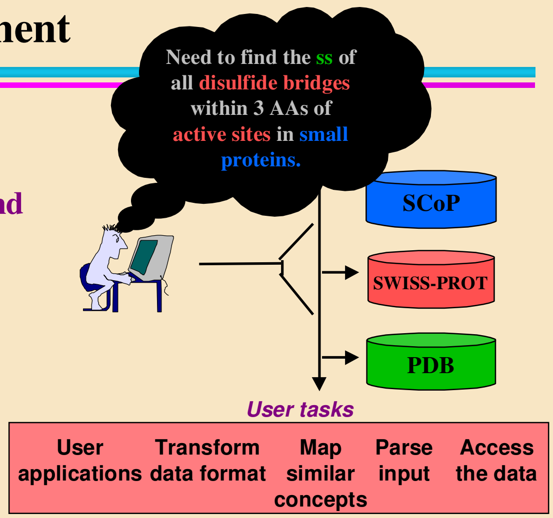
Current Environment

- Sources are
  - Distributed
  - Autonomous
  - Heterogeneous
  - Conflicting



Current Environment

- Data is
  - Hard to find
  - Hard to understand
  - Hard to reconcile
  - Hard to analyze

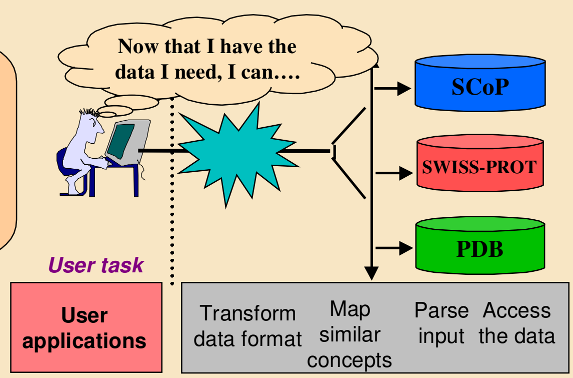


Scientists waste time and energy doing data management.

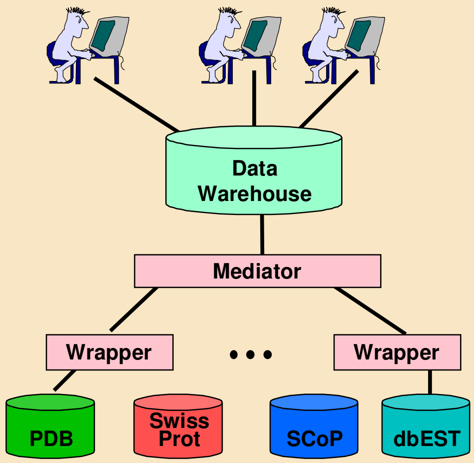
What is our ideal environment?

A *single* location that provides *effective* access to a *consistent* view of data from *many* sources through an *intuitive and useful* interface.

Businesses use data warehouses to accomplish this.

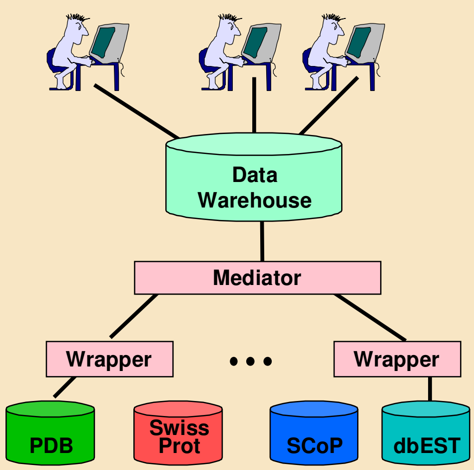


Data warehouses



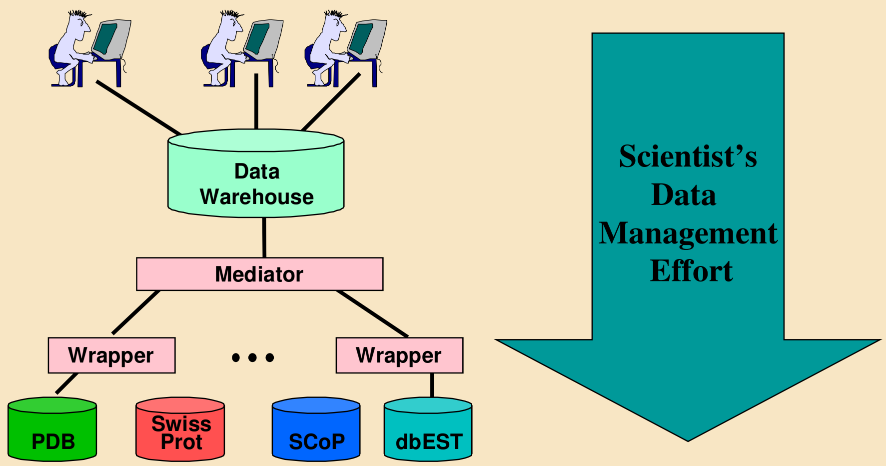
A data warehouse is a repository that provides a single access point to a collection of data obtained from a set of distributed, heterogeneous sources.

Data warehouses



- Interfaces
  - provide intuitive access to the data
  - possibly change data format to meet user expectations
- Warehouse
  - stores a consistent view of data in a local repository
- Mediator
  - transform data from source format to warehouse format
- Wrappers
  - read data from source into internal representation

Data warehouses



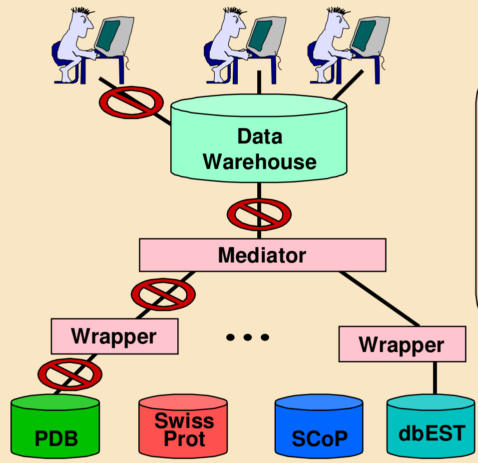
Warehouses are complex!

- Resolve semantic and syntactic conflicts.
- Define a global schema.
- Write:
  - Wrappers.
  - Mediators.
  - Interfaces.

In business, warehouses:
- take 2yrs +
- cost $5 million +

These are expensive, and time consuming ventures.

Warehouses don’t work in dynamic domains



When schemata are modified,
or new sources are added,
wrappers and mediators break.

Scientific domains are highly dynamic

Changes in scientific warehouses occur because:

- Our understanding of the data is constantly evolving.
- Constantly need more information.
- Funding sources change.

Even though they work in business, existing data warehousing technology is too expensive to maintain in dynamic, scientific environments.

Example 1

Adding a new data source, from a new domain, into an existing warehouse.
Users want access to the new data ASAP.

How does the new data relate to what has already been integrated?

Warehouse (and mediator) grow in complexity!

well-defined semantics & structured mediator

Example 2

Building a warehouse combining data from 50 sources, each of which changes once a year.
On average, that is one change EVERY week.

What happens if it takes longer than a week to update the warehouse?

The warehouse is immediately out of date!

time to update < time between changes

Other approaches for accessing data

- Multi-database
- Federated database
- Warehouse integrating only high level concepts

Other approaches for accessing data

~~Multi-database~~

~~Federated database~~

~~High-level warehouse~~

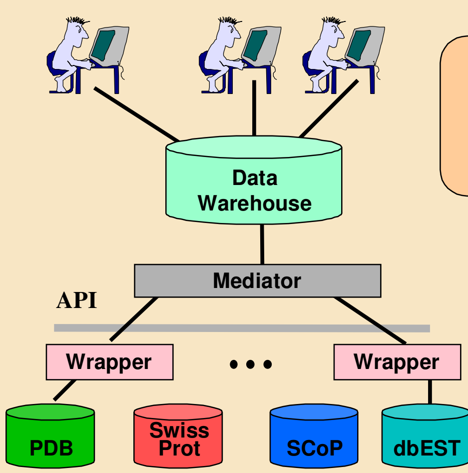
Nothing seems to work!

A *single* location that provides *effective* access to a *consistent* view of data from *many* (dynamic) sources through an *intuitive and useful* interface.

Data warehouses meet 4/5 of these requirements.

Solution:
Reduce the cost of building and maintaining the data warehouse.

Key Insights



Need a comprehensive approach to managing change.

Extensive use of meta-data can dramatically reduce maintenance costs.

Results

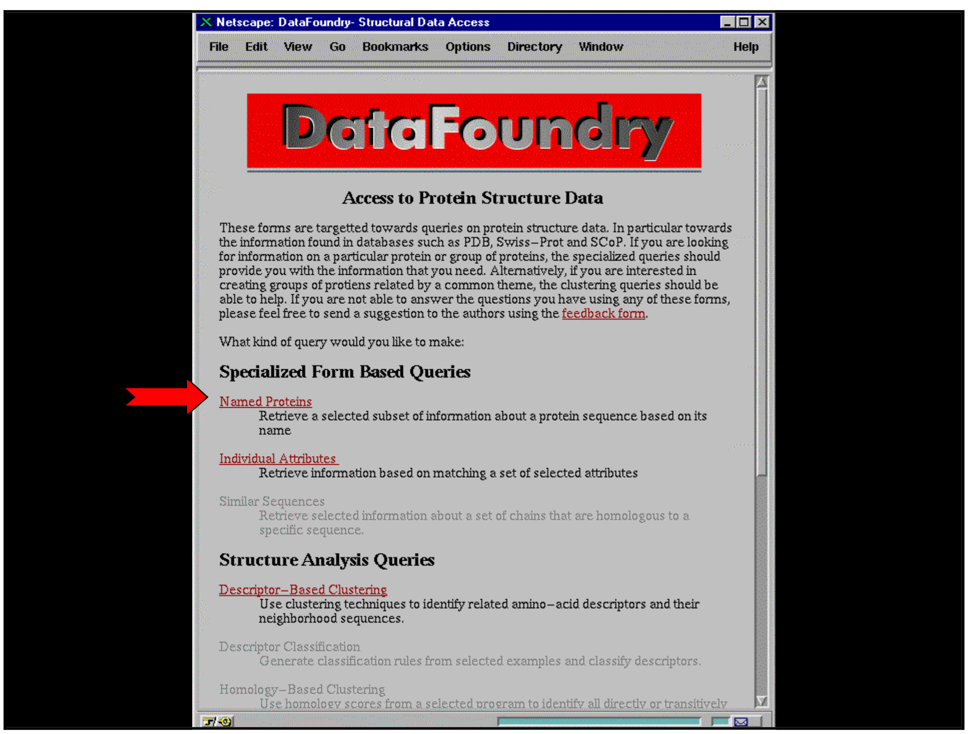
Integrating SCoP into warehouse that already contains PDB and SWISS-PROT.

Activity/ integration style	manual	meta-data	diff	%diff
understanding SCOP	2.0	2.0	0.0	0
writing wrapper	4.5	2.5	2.0	44%
modifying schema	0.5	0.5	0.0	0
writing mediator	4.0	0.0	4.0	---
modifying meta-data	0.0	1.0	(1.0)	---
total time in days	11.0	6.0	5.0	45%

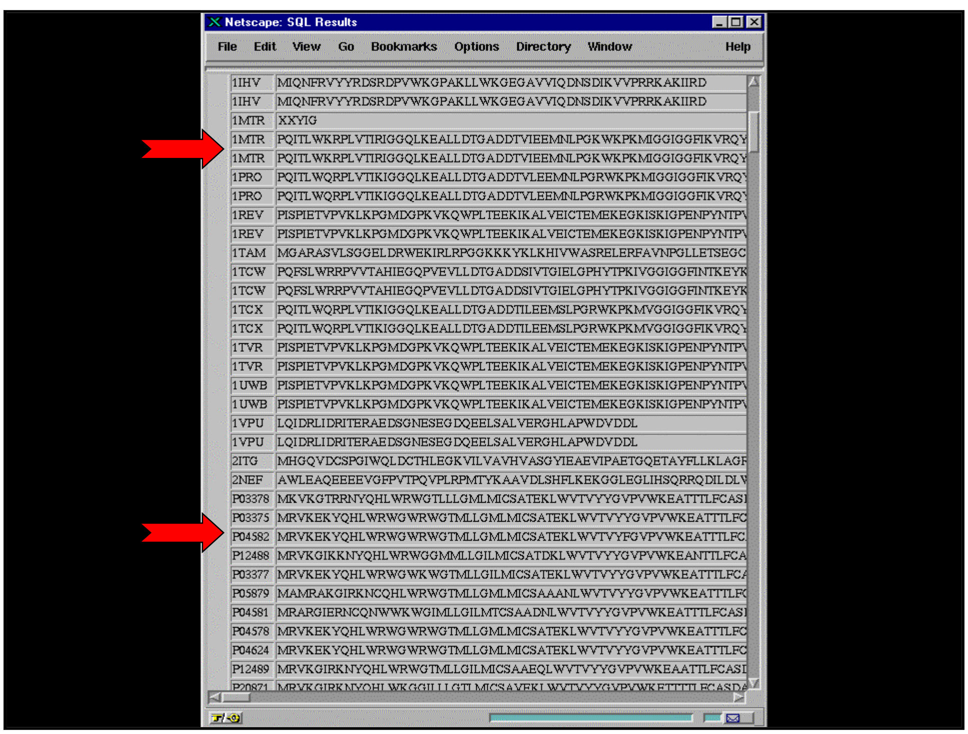
DataFoundry Prototype

- 3 data sources
  - PDB, SWISS-PROT, SCoP
- 7Gb Sybase database
  - 13k structures
  - 13k taxonomy mappings
  - 88k sequences
  - 100k+ authors
  - 220k structural classifications
  - 48MM atoms
- Novice, intermediate and expert interfaces


Netscape: DataFoundry - Database Access Home
File Edit View Go Bookmarks Options Directory Window Help
DataFoundry
Welcome to the DataFoundry Data Warehouse Page
We have developed several formed–based interfaces to our warehouse in order to provide you with the best service possible. In addition to domain specific queries, certain users can create their own, ad hoc queries against the database. If you feel you need this level of access, please contact the authors directly.
Note that before you can access the warehouse, you must Register.
What type of information are you primarily interested in:
Specific Domain Queries
Structural Biology
Sequencing
Functional Genomics
Publications
Free Form Queries
Free–form SQL queries (internal only)
Graphical query interface
General Warehouse Information
Browse the ontology
Browse the email archive
Provide feedback to the authors
OVERVIEW
TEAM
C.S. RESEARCH
SUPERCOMPUTING
ALTERNATIVE USES
PUBLICATIONS
GENE FOUNDRY
RELATED TOPICS

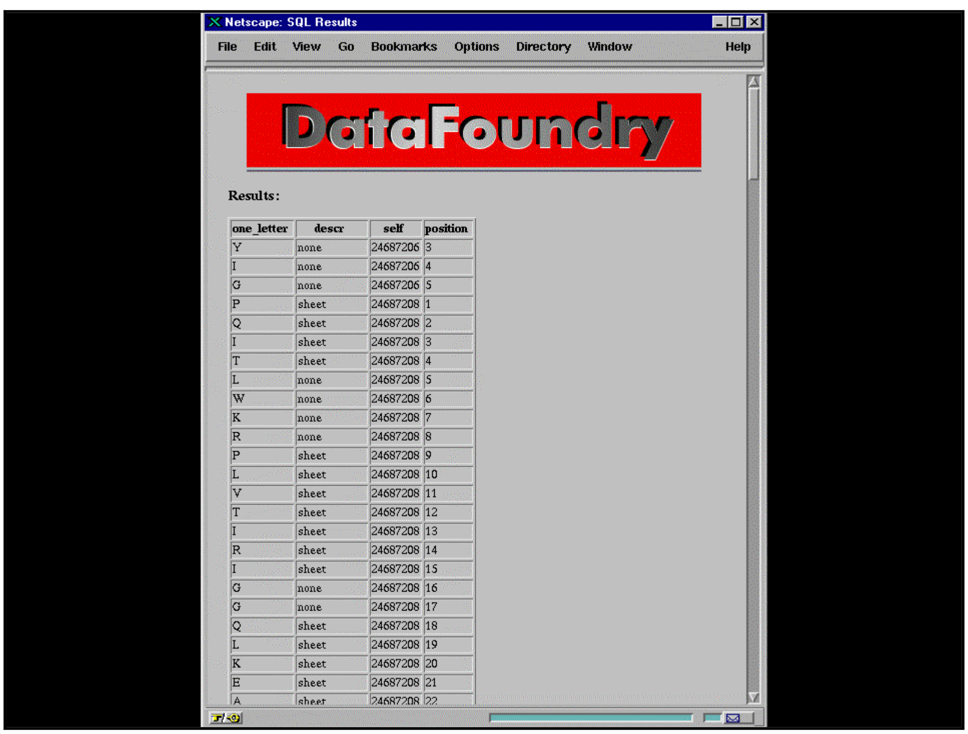

Netscape: DataFoundry- Structural Data Access
File Edit View Go Bookmarks Options Directory Window Help
DataFoundry
Access to Protein Structure Data
These forms are targetted towards queries on protein structure data. In particular towards the information found in databases such as PDB, Swiss–Prot and SCoP. If you are looking for information on a particular protein or group of proteins, the specialized queries should provide you with the information that you need. Alternatively, if you are interested in creating groups of protiens related by a common theme, the clustering queries should be able to help. If you are not able to answer the questions you have using any of these forms, please feel free to send a suggestion to the authors using the feedback form.
What kind of query would you like to make:
Specialized Form Based Queries
Named Proteins
Retrieve a selected subset of information about a protein sequence based on its name
Individual Attributes
Retrieve information based on matching a set of selected attributes
Similar Sequences
Retrieve selected information about a set of chains that are homologous to a specific sequence.
Structure Analysis Queries
Descriptor–Based Clustering
Use clustering techniques to identify related amino–acid descriptors and their neighborhood sequences.
Descriptor Classification
Generate classification rules from selected examples and classify descriptors.
Homology–Based Clustering

Netscape: DataFoundry - Sequence Name Search
File Edit View Go Bookmarks Options Directory Window Help
DataFoundry
Name Based Protein Search
This form will return the selected information about a set of protiens identified by name. Valid names include PDB files, Swiss–Prot accession numbers, keywords etc.
Protein name: AIDS
OR
Protein list:
Select protein attributes to be returned:
Name: Sequences: Author:
Resolution: Source: R–Value:
Articles: Secondary Structure: Length:
Coordinates: Heterogens: Comments:
Initial Date: Modified Date: Pub. Author:
Select residue attributes to be returned:
One–letter name: Three–letter name: Position:
Structure (text): Structure (numeric): Hydrophobicity:
Submit Query Reset

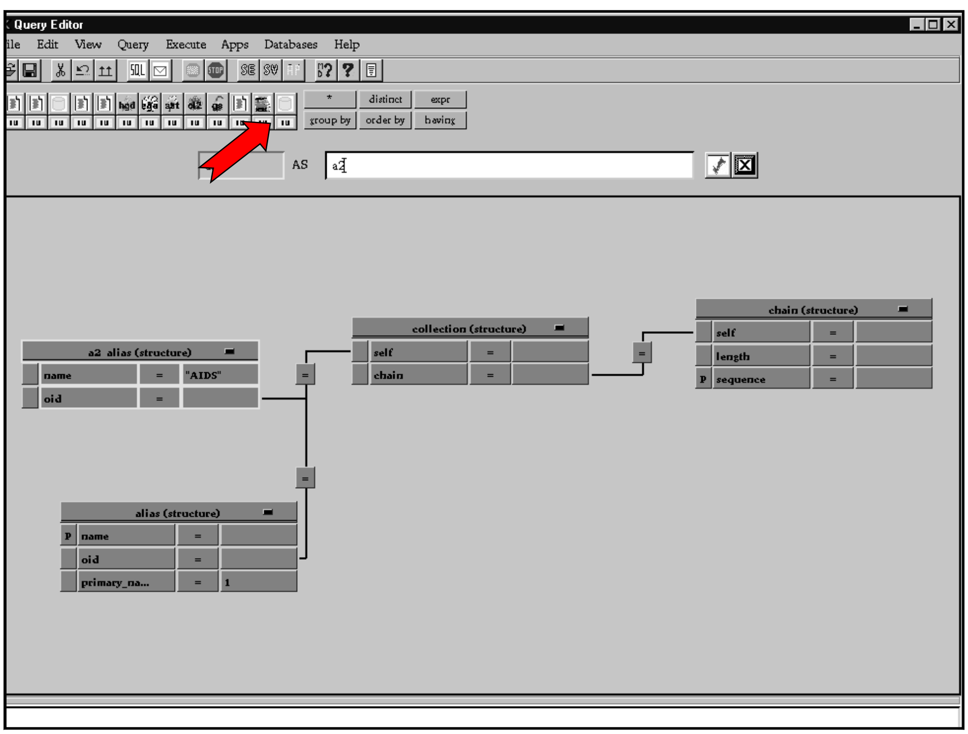


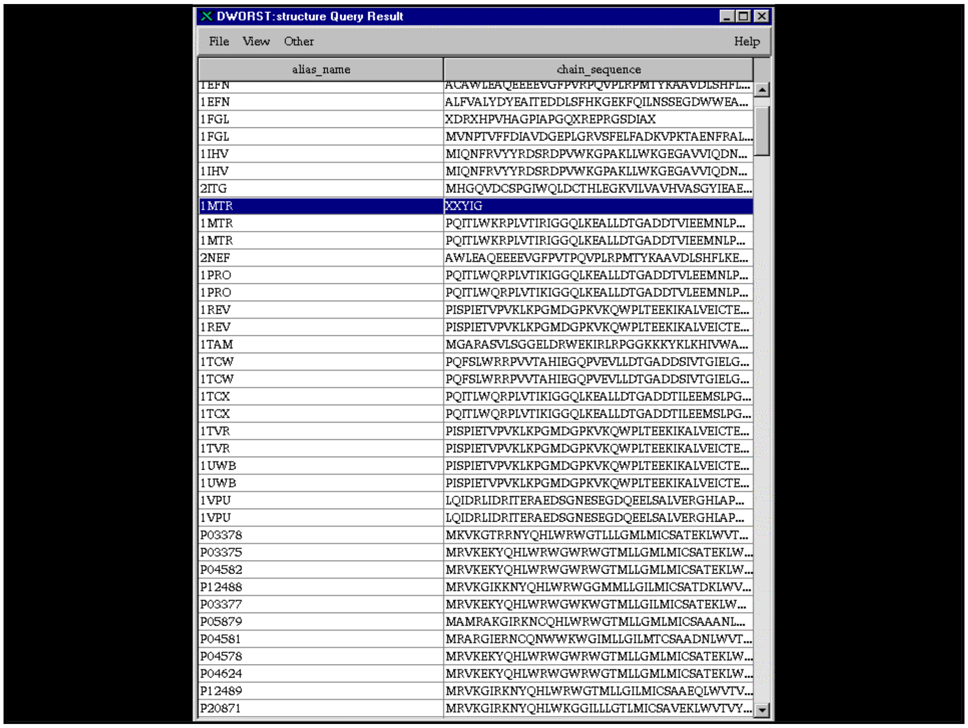
Netscape: SQL Results
File Edit View Go Bookmarks Options Directory Window Help
1IHV MIQNFRVYYRDSRDPVWKGPAKLLWKGEGAVVIQDNSDIKVVPRRKAKIIRD
1IHV MIQNFRVYYRDSRDPVWKGPAKLLWKGEGAVVIQDNSDIKVVPRRKAKIIRD
1MTR XXYIG
1MTR PQITLWKRPLVTIRIGGQLKEALLDTGADDTVIEEMNLPGKWKPKMIGGIGGFIKVRQ
1MTR PQITLWKRPLVTIRIGGQLKEALLDTGADDTVIEEMNLPGKWKPKMIGGIGGFIKVRQ
1PRO PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGGIGGFIKVRQ
1PRO PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGGIGGFIKVRQ
1REV PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTEMEKEGKISKIGPENPYNTP
1REV PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTEMEKEGKISKIGPENPYNTP
1TAM MGARASVLSGGELDRWEKIRLRPGGKKKYKLKHIVWASRELERFAVNPGLLETSEGC
1TCW PQFSLWRRPVVTAHIEGQPVEVLLDTGADDSIVTGIELGPHYTPKIVGGIGGFINTKEYK
1TCW PQFSLWRRPVVTAHIEGQPVEVLLDTGADDSIVTGIELGPHYTPKIVGGIGGFINTKEYK
1TCX PQITLWQRPLVTIKIGGQLKEALLDTGADDTILEEMSLPGRWKPKMVGGIGGFIKVRQ
1TCX PQITLWQRPLVTIKIGGQLKEALLDTGADDTILEEMSLPGRWKPKMVGGIGGFIKVRQ
1TVR PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTEMEKEGKISKIGPENPYNTP
1TVR PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTEMEKEGKISKIGPENPYNTP
1UWB PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTEMEKEGKISKIGPENPYNTP
1UWB PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTEMEKEGKISKIGPENPYNTP
1VPU LQIDRLIDRITERAEDSGNESEGDQEELSALVERGHLAPWDVDDL
1VPU LQIDRLIDRITERAEDSGNESEGDQEELSALVERGHLAPWDVDDL
2ITG MHGQVDCSPGIWQLDCTHLEGKVILVAVHVASGYIEAEVIPAETGQETAYFLLKLAG
2NEF AWLEAQEEEEVGFPVTPQVPLRPMTYKAAVDLSHFLKEKGGLEGLIHSQRRQDILDL
P03378 MKVKGTRRNYQHLWRWGTLLLGMLMICSATEKLWVTVYYGVPVWKEATTTLFCAS
P03375 MRVKEKYQHLWRWGWRWGTMLLGMLMICSATEKLWVTVYYGVPVWKEATTTLFC
P04582 MRVKEKYQHLWRWGWRWGTMLLGMLMICSATEKLWVTVYFGVPVWKEATTTLFC
P12488 MRVKGIKKNYQHLWRWGGMMLLGILMICSATDKLWVTVYYGVPVWKEANTTLFCA
P03377 MRVKEKYQHLWRWGWKWGTMLLGILMICSATEKLWVTVYYGVPVWKEATTTLFC
P05879 MAMRAKGIRKNCQHLWRWGTMLLGMLMICSAAANLWVTVYYGVPVWKEATTTLF
P04581 MRARGIERNCQNWWKWGIMLLGILMTCSAADNLWVTVYYGVPVWKEATTTLFCAS
P04578 MRVKEKYQHLWRWGWRWGTMLLGMLMICSATEKLWVTVYYGVPVWKEATTTLFC
P04624 MRVKEKYQHLWRWGWRWGTMLLGMLMICSATEKLWVTVYYGVPVWKEATTTLFC
P12489 MRVKGIRKNYQHLWRWGTMLLGILMICSAAEQLWVTVYYGVPVWKEAATTLFCASI


Netscape: DataFoundry - Sequence Name Search
File Edit View Go Bookmarks Options Directory Window Help
DataFoundry
Name Based Protein Search
This form will return the selected information about a set of protiens identified by name. Valid names include PDB files, Swiss–Prot accession numbers, keywords etc.
Protein name: 1MTR
OR
Protein list:
Select protein attributes to be returned:
Name: Sequences: Author:
Resolution: Source: R–Value:
Articles: Secondary Structure: Length:
Coordinates: Heterogens: Comments:
Initial Date: Modified Date: Pub. Author:
Select residue attributes to be returned:
One–letter name: Three–letter name: Position:
Structure (text): Structure (numeric): Hydrophobicity:
Submit Query Reset


Netscape: SQL Results
File Edit View Go Bookmarks Options Directory Window Help
DataFoundry
Results:
one_letter descr self position
Y none 24687206 3
I none 24687206 4
G none 24687206 5
P sheet 24687208 1
Q sheet 24687208 2
I sheet 24687208 3
T sheet 24687208 4
L none 24687208 5
W none 24687208 6
K none 24687208 7
R none 24687208 8
P sheet 24687208 9
L sheet 24687208 10
V sheet 24687208 11
T sheet 24687208 12
I sheet 24687208 13
R sheet 24687208 14
I sheet 24687208 15
G none 24687208 16
G none 24687208 17
Q sheet 24687208 18
L sheet 24687208 19
K sheet 24687208 20
E sheet 24687208 21
A sheet 24687208 22


Netscape: DataFoundry - Database Access Home
File Edit View Go Bookmarks Options Directory Window Help
DataFoundry
Welcome to the DataFoundry Data Warehouse Page
We have developed several formed–based interfaces to our warehouse in order to provide you with the best service possible. In addition to domain specific queries, certain users can create their own, ad hoc queries against the database. If you feel you need this level of access, please contact the authors directly.
Note that before you can access the warehouse, you must Register.
What type of information are you primarily interested in:
Specific Domain Queries
Structural Biology
Sequencing
Functional Genomics
Publications
Free Form Queries
Free–form SQL queries (internal only)
Graphical query interface
General Warehouse Information
Browse the ontology
Browse the email archive
Provide feedback to the authors
OVERVIEW
TEAM
C.S. RESEARCH
SUPERCOMPUTING
ALTERNATIVE USES
PUBLICATIONS
GENE FOUNDRY
RELATED TOPICS


Query Editor
ile Edit View Query Execute Apps Databases Help
distinct
expr
group by
order by
having
AS
a2
a2 alias (structure)
name
=
"AIDS"
oid
collection (structure)
self
chain
chain (structure)
length
sequence
alias (structure)
primary_na...
1



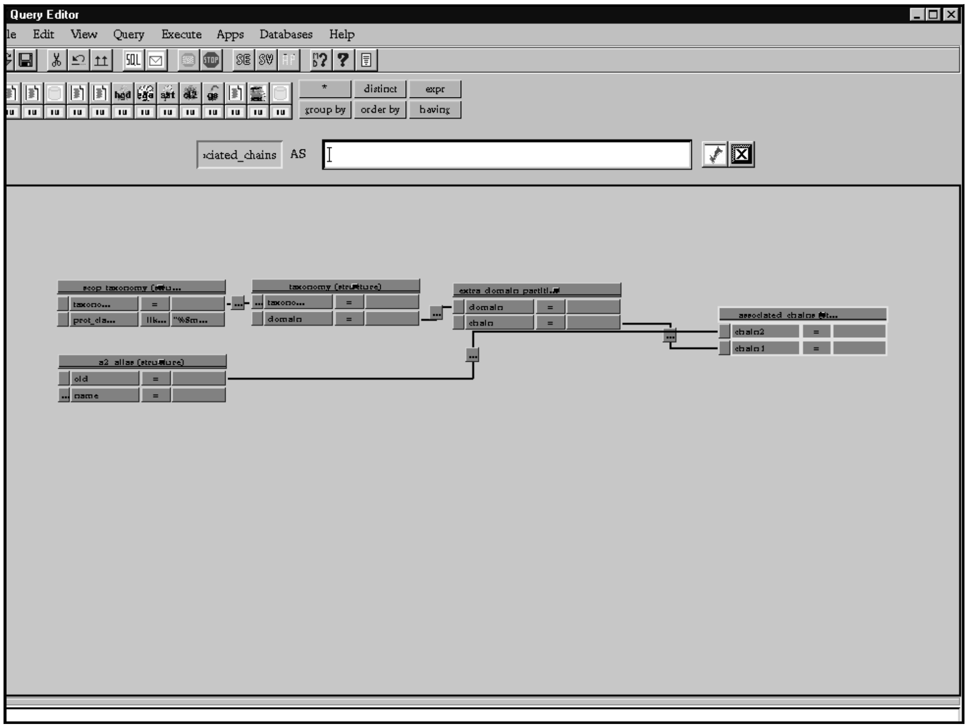
DWORST:structure Query Result

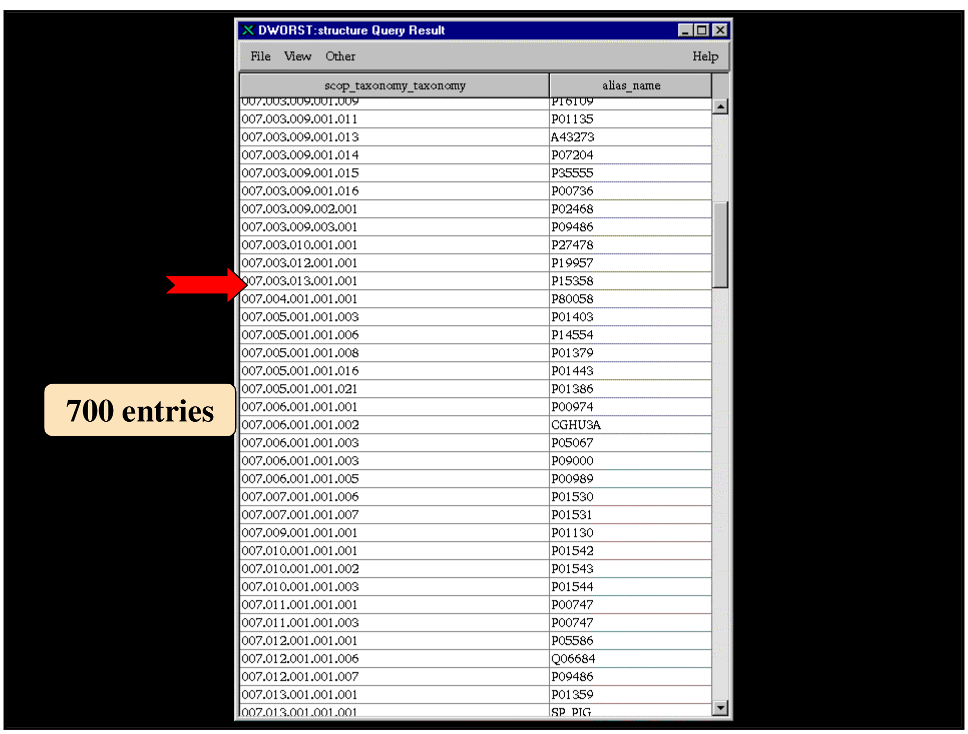
File View Other Help

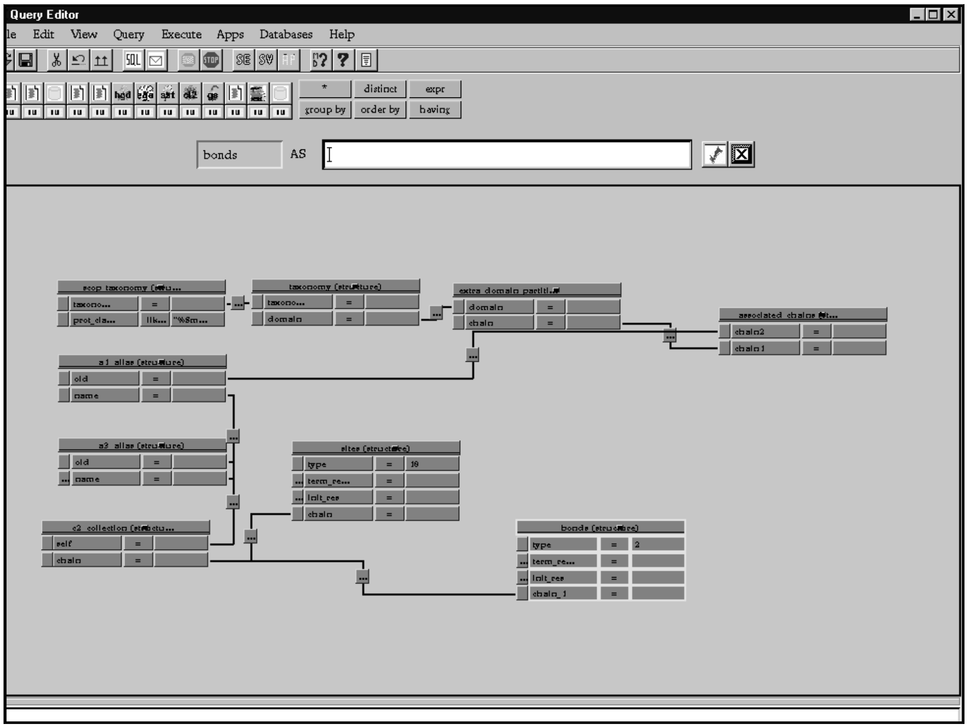
alias_name	chain_sequence
1EFN	ACAWLEAQEEEEVGFPVRPQVPLRPMTYKAAVDLSHFL...
1EFN	ALFVALYDYEAITEDDLSFHKGEKFQILNSSEGDWWEA...
1FGL	XDRXHPVHAGPIAPGQXREPRGSDIAX
1FGL	MVNPTVFFDIAVDGEPLGRVSFELFADKVPKTAENFRAL...
1IHV	MIQNFRVYYRDSRDPVWKGPAKLLWKGEGAVVIQDN...
1IHV	MIQNFRVYYRDSRDPVWKGPAKLLWKGEGAVVIQDN...
2ITG	MHGQVDCSPGIWQLDCTHLEGKVILVAVHVASGYIEAE...
1MTR	XXYIG
1MTR	PQITLWKRPLVTIRIGGQLKEALLDTGADDTVIEEMNLP...
1MTR	PQITLWKRPLVTIRIGGQLKEALLDTGADDTVIEEMNLP...
2NEF	AWLEAQEEEEVGFPVTPQVPLRPMTYKAAVDLSHFLKE...
1PRO	PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLP...
1PRO	PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLP...
1REV	PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTE...
1REV	PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTE...
1TAM	MGARASVLSGGELDRWEKIRLRPGGKKKYKLKHIVWA...
1TCW	PQFSLWRRPVVTAHIEGQPVEVLLDTGADDSIVTGIELG...
1TCW	PQFSLWRRPVVTAHIEGQPVEVLLDTGADDSIVTGIELG...
1TCX	PQITLWQRPLVTIKIGGQLKEALLDTGADDTILEEMSLPG...
1TCX	PQITLWQRPLVTIKIGGQLKEALLDTGADDTILEEMSLPG...
1TVR	PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTE...
1TVR	PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTE...
1UWB	PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTE...
1UWB	PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTE...
1VPU	LQIDRLIDRITERAEDSGNESEGDQEELSALVERGHLAP...
1VPU	LQIDRLIDRITERAEDSGNESEGDQEELSALVERGHLAP...
P03378	MKVKGTRRNYQHLWRWGTLLLGMLMICSATEKLWVT...
P03375	MRVKEKYQHLWRWGWRWGTMLLGMLMICSATEKLW...
P04582	MRVKEKYQHLWRWGWRWGTMLLGMLMICSATEKLW...
P12488	MRVKGIKKNYQHLWRWGGMMLLGILMICSATDKLWV...
P03377	MRVKEKYQHLWRWGWKWGTMLLGILMICSATEKLW...
P05879	MAMRAKGIRKNCQHLWRWGTMLLGMLMICSAAANL...
P04581	MRARGIERNCQNWWKWGIMLLGILMTCSAADNLWVT...
P04578	MRVKEKYQHLWRWGWRWGTMLLGMLMICSATEKLW...
P04624	MRVKEKYQHLWRWGWRWGTMLLGMLMICSATEKLW...
P12489	MRVKGIRKNYQHLWRWGTMLLGILMICSAAEQLWVTV...
P20871	MRVKGIRKNYQHLWKGGILLLGTLMICSAVEKLWVTVY...

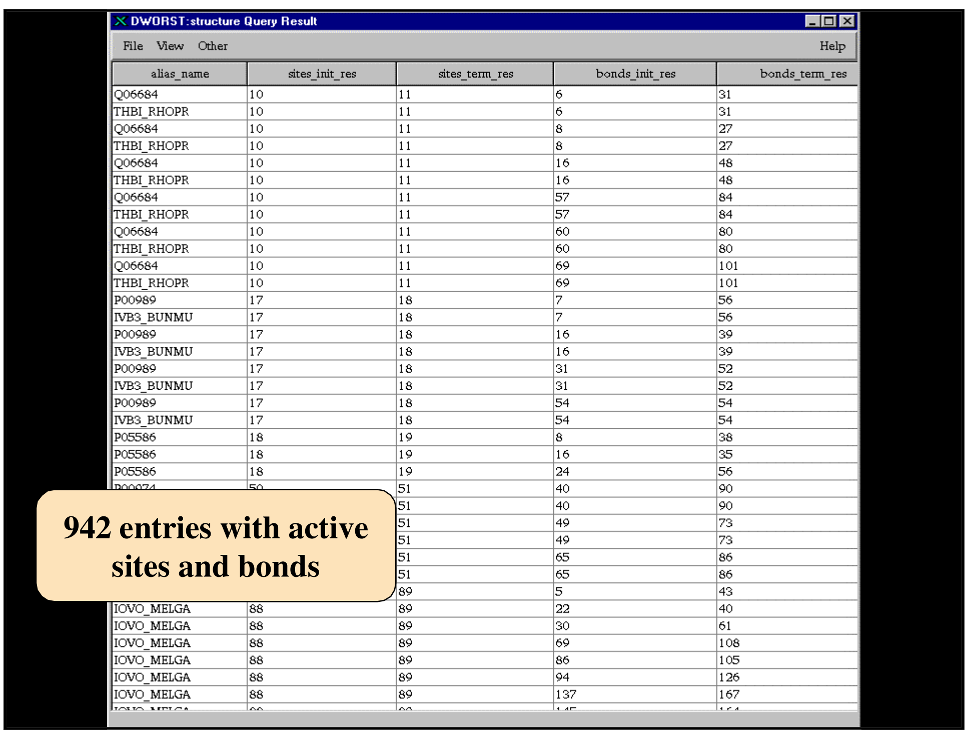
DataFoundry Prototype

CASC

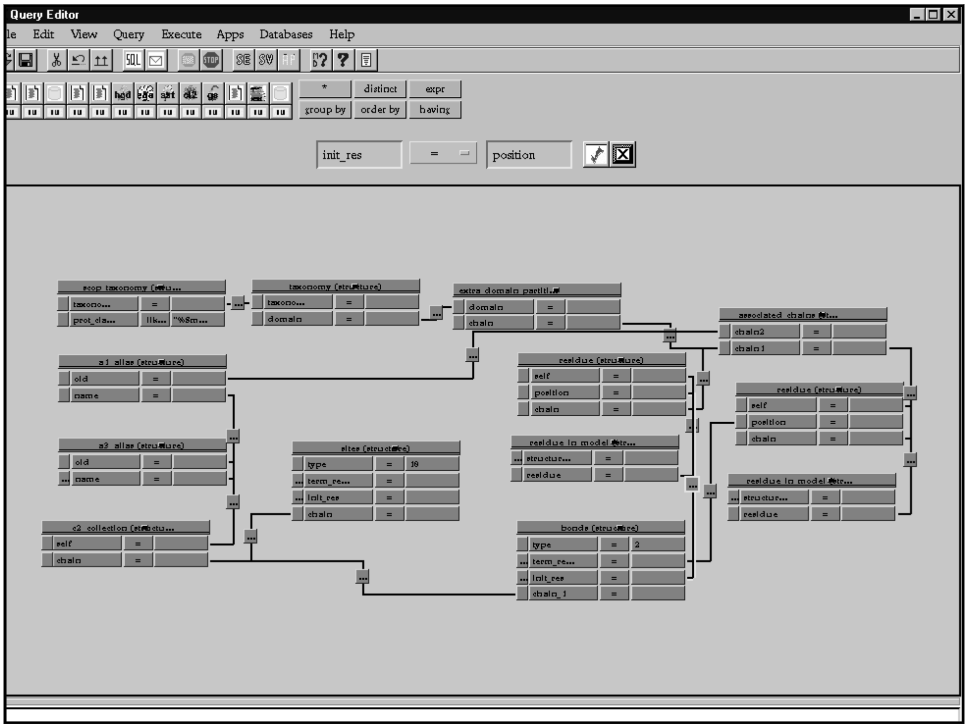

Query Editor
le Edit View Query Execute Apps Databases Help
*
distinct
expr
group by
order by
having
ciated_chains AS

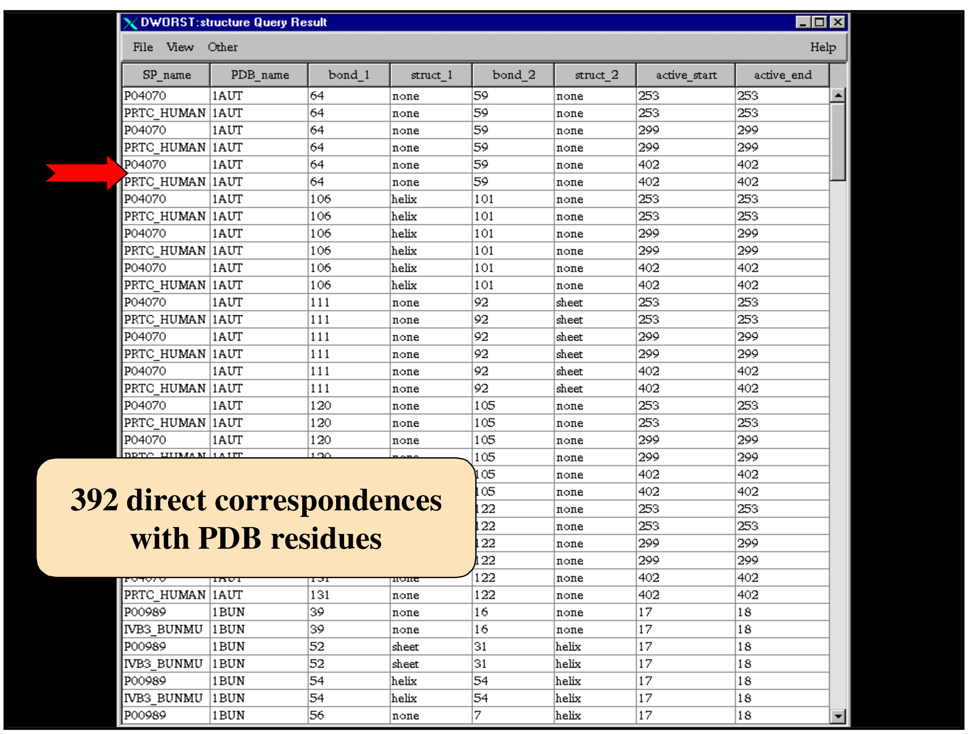

DWORST:structure Query Result
File View Other Help
scop_taxonomy_taxonomy alias_name
007.003.009.001.011 P01135
007.003.009.001.013 A43273
007.003.009.001.014 P07204
007.003.009.001.015 P35555
007.003.009.001.016 P00736
007.003.009.002.001 P02468
007.003.009.003.001 P09486
007.003.010.001.001 P27478
007.003.012.001.001 P19957
07.003.013.001.001 P15358
007.004.001.001.001 P80058
007.005.001.001.003 P01403
007.005.001.001.006 P14554
007.005.001.001.008 P01379
007.005.001.001.016 P01443
007.005.001.001.021 P01386
007.006.001.001.001 P00974
007.006.001.001.002 CGHU3A
007.006.001.001.003 P05067
007.006.001.001.003 P09000
007.006.001.001.005 P00989
007.007.001.001.006 P01530
007.007.001.001.007 P01531
007.009.001.001.001 P01130
007.010.001.001.001 P01542
007.010.001.001.002 P01543
007.010.001.001.003 P01544
007.011.001.001.001 P00747
007.011.001.001.003 P00747
007.012.001.001.001 P05586
007.012.001.001.006 Q06684
007.012.001.001.007 P09486
007.013.001.001.001 P01359
700 entries



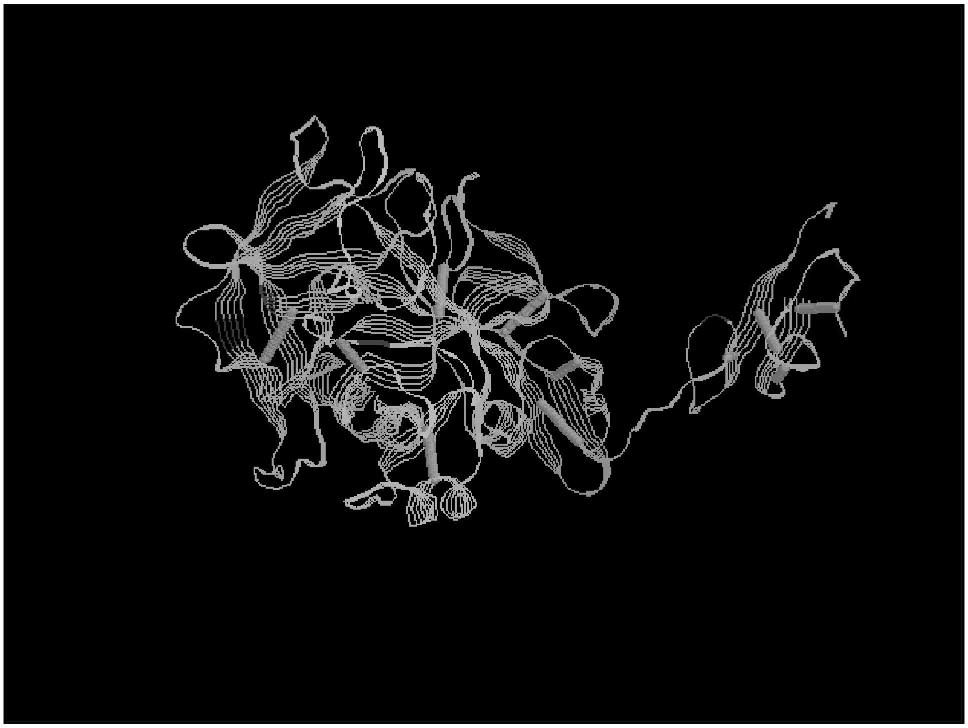


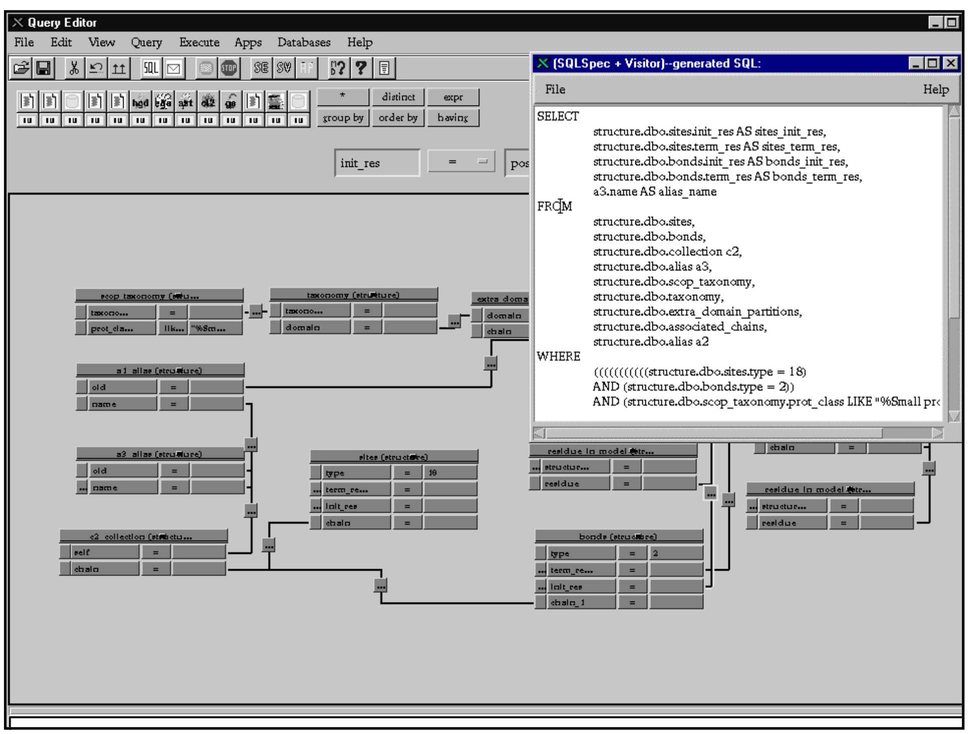
alias_name	sites_init_res	sites_term_res	bonds_init_res	bonds_term_res
Q06684	10	11	6	31
THBI_RHOPR	10	11	6	31
Q06684	10	11	8	27
THBI_RHOPR	10	11	8	27
Q06684	10	11	16	48
THBI_RHOPR	10	11	16	48
Q06684	10	11	57	84
THBI_RHOPR	10	11	57	84
Q06684	10	11	60	80
THBI_RHOPR	10	11	60	80
Q06684	10	11	69	101
THBI_RHOPR	10	11	69	101
P00989	17	18	7	56
IVB3_BUNMU	17	18	7	56
P00989	17	18	16	39
IVB3_BUNMU	17	18	16	39
P00989	17	18	31	52
IVB3_BUNMU	17	18	31	52
P00989	17	18	54	54
IVB3_BUNMU	17	18	54	54
P05586	18	19	8	38
P05586	18	19	16	35
P05586	18	19	24	56
P00974	50	51	40	90
[illegible]	[illegible]	51	40	90
[illegible]	[illegible]	51	49	73
[illegible]	[illegible]	51	49	73
[illegible]	[illegible]	51	65	86
[illegible]	[illegible]	51	65	86
[illegible]	[illegible]	89	5	43
IOVO_MELGA	88	89	22	40
IOVO_MELGA	88	89	30	61
IOVO_MELGA	88	89	69	108
IOVO_MELGA	88	89	86	105
IOVO_MELGA	88	89	94	126
IOVO_MELGA	88	89	137	167





SP_name	PDB_name	bond_1	struct_1	bond_2	struct_2	active_start	active_end
P04070	1AUT	64	none	59	none	253	253
PRTC_HUMAN	1AUT	64	none	59	none	253	253
P04070	1AUT	64	none	59	none	299	299
PRTC_HUMAN	1AUT	64	none	59	none	299	299
P04070	1AUT	64	none	59	none	402	402
PRTC_HUMAN	1AUT	64	none	59	none	402	402
P04070	1AUT	106	helix	101	none	253	253
PRTC_HUMAN	1AUT	106	helix	101	none	253	253
P04070	1AUT	106	helix	101	none	299	299
PRTC_HUMAN	1AUT	106	helix	101	none	299	299
P04070	1AUT	106	helix	101	none	402	402
PRTC_HUMAN	1AUT	106	helix	101	none	402	402
P04070	1AUT	111	none	92	sheet	253	253
PRTC_HUMAN	1AUT	111	none	92	sheet	253	253
P04070	1AUT	111	none	92	sheet	299	299
PRTC_HUMAN	1AUT	111	none	92	sheet	299	299
P04070	1AUT	111	none	92	sheet	402	402
PRTC_HUMAN	1AUT	111	none	92	sheet	402	402
P04070	1AUT	120	none	105	none	253	253
PRTC_HUMAN	1AUT	120	none	105	none	253	253
P04070	1AUT	120	none	105	none	299	299
[illegible]	[illegible]	[illegible]	[illegible]	105	none	299	299
[illegible]	[illegible]	[illegible]	[illegible]	105	none	402	402
[illegible]	[illegible]	[illegible]	[illegible]	105	none	402	402
[illegible]	[illegible]	[illegible]	[illegible]	122	none	253	253
[illegible]	[illegible]	[illegible]	[illegible]	122	none	253	253
[illegible]	[illegible]	[illegible]	[illegible]	122	none	299	299
[illegible]	[illegible]	[illegible]	[illegible]	122	none	299	299
[illegible]	[illegible]	[illegible]	[illegible]	122	none	402	402
PRTC_HUMAN	1AUT	131	none	122	none	402	402
P00989	1BUN	39	none	16	none	17	18
IVB3_BUNMU	1BUN	39	none	16	none	17	18
P00989	1BUN	52	sheet	31	helix	17	18
IVB3_BUNMU	1BUN	52	sheet	31	helix	17	18
P00989	1BUN	54	helix	54	helix	17	18
IVB3_BUNMU	1BUN	54	helix	54	helix	17	18
P00989	1BUN	56	none	7	helix	17	18




Query Editor
File Edit View Query Execute Apps Databases Help
distinct
expr
group by
order by
having
init_res
(SQLSpec + Visitor)--generated SQL:
File
Help
SELECT
structure.dbo.sites.init_res AS sites_init_res,
structure.dbo.sites.term_res AS sites_term_res,
structure.dbo.bonds.init_res AS bonds_init_res,
structure.dbo.bonds.term_res AS bonds_term_res,
a3.name AS alias_name
FROM
structure.dbo.sites,
structure.dbo.bonds,
structure.dbo.collection c2,
structure.dbo.alias a3,
structure.dbo.scop_taxonomy,
structure.dbo.taxonomy,
structure.dbo.extra_domain_partitions,
structure.dbo.associated_chains,
structure.dbo.alias a2
WHERE
((((((((((structure.dbo.sites.type = 18)
AND (structure.dbo.bonds.type = 2))
AND (structure.dbo.scop_taxonomy.prot_class LIKE "%Small pr

Conclusions

Data warehouses could provide a fast, reliable, consistent view of data from multiple sources, if their maintenance costs were substantially reduced.

We are successfully addressing this challenge.

Work performed under the auspices of the U.S. DOE by LLNL under contract No. W-7405-ENG-48